Editorial

Andreas Heuer*

Research Data Management

https://doi.org/10.1515/itit-2020-0002

1 Introduction to research data management

More and more areas in science, even systematic and theory-based sciences such as Mathematics, are evolving into *data-driven sciences* where a lot of data are used or produced to support the research work. These can result from measurements (from experiments in labs or more recently from always-on-sensors such as microphones and cameras in the Internet of Things) or from modelling and simulation processes. A lot of research areas from natural sciences, medical sciences, engineering, among others, are more and more data-driven. For these, research data management is becoming a crucial issue.

A completely different area are the less measurement- and sensor-data-driven humanities research areas (Digital Humanities), in which work is very text- or document-centered. We can call these document-driven sciences. Also completely different are sciences relying on non-digital artifacts such as soil, water, or material samples being first class research objects that have to be "stored" to be able to reproduce research results afterwards. We can call these artifact-based sciences. Both document-driven and artifact-based sciences will also rely on digital infrastructures for digital documents and scanned historical texts, as well as on digital infrastructures for secondary (derived) data from experiments with non-digital artifacts and metadata for these experiments and artifacts. Research data management is therefore also a crucial issue for these types of sciences.

The articles in this special issue will deal more with data-driven scenarios, especially with research data in Mathematics and in interdisciplinary research centres between Biomedicine, Electrical Engineering, and Computer Science.

Research Data Management aims at gathering, capturing, storing, tracking, and archiving all the data being produced in scientific projects and experiments. Besides these data, all the processing steps on these data – eventually resulting in scientific publications – have to be stored as well.

Many conferences and workshops are dedicated to this topic and research funders almost always expect concepts regarding sustainability, traceability, and transparent presentation and publication of research data. Therefore, universities must increasingly develop organizational concepts for research data management and implement pragmatic solutions for research data management in a timely manner.

In the context of Open Science [8] and the European Open Science Cloud (EOSC), the FAIR principles seem to become a common and widely accepted conceptual basis for future research data infrastructures: data must be Findable, Accessible, Interoperable, and Reusable in order to advance the discoverability, reuse, and reproducibility of research results [11].

If we only look into the aspect of reusability and reproducibility in more detail, one can find a lot of different notions and quality criteria for this principle in research and practice. The ACM Task Force on Data, Software and Reproducibility [4] distinguishes between repeatability, replicability, and reproducibility (among others) of research results (see also [6]). In Table 1, the differences between these notions are depicted. While for repeatability, the same research group publishing the original research result should be able to do the same evaluation yielding the same result, for replicability and reproducibility another research group (or simply a reviewer) should be able to do the evaluation resulting in the same output. While for repeatability and replicability, the same evaluation technique as in the original experiment should be applied to prove the same result, for reproducibility the reviewer or

Table 1: A closer look into the "R" of FAIR: From Repeatability to Reproducibility.

Aim	Who?	How?
Repeatability	same group	same evaluation
Replicability	different group	same evaluation
Reproducibility	different group	different evaluation

***Corresponding author: Andreas Heuer,** University of Rostock, Computer Science Institute, 18059 Rostock, Germany, e-mail: andreas.heuer@uni-rostock.de, ORCID: https://orcid.org/0000-0002-6163-6649

control research group should produce the same result even applying a different evaluation technique.

As one can see, for replicability and reproducibility, the FAIR principles of accessibility and interoperability are a prerequisite for a different research group to be able to check the original results, but it is also important to have accessibility and interoperability not only for the data, but also for the software tools used including the evaluation scripts performed. One of the articles in this special issue will extend the FAIR principle from research data to research software manipulating or evaluating the research data.

Research data management is not only a scientific discipline in Computer Science. Universities and research institutes have to provide organizational structures and processes and pragmatic solutions (hardware and software resources) to implement first, simple tasks of research data management. Though a lot of research data management platforms are already available [1], they focus on the support of only a subset of the tasks in the scientific workflow, mainly in the publication and archiving phase at the end of the research project. One of the articles in this special issue will extend a classical data sharing platform to be able to support the research work in interdisciplinary teams more directly.

2 Scope of this special issue

This special issue asked for contributions from all areas of research data management, including organizational, pragmatic, and research aspects, as well as

- integrating Linked Open Data,
- privacy aspects, copyright and licenses, data curation, and archiving,
- approaches to make research data Findable, Accessible, Interoperable, and Reusable,
- support of replicable, reproducible, sustainable, explainable, and transparent research, as well as
- database research topics applicable to research data management such as temporal databases, data integration, schema evolution, and provenance management.

In the database research community, there is already a significant amount of work in the areas of data and (scientific) workflow provenance [7] adapting W3C standards [9] to scientific processes [10], as well as trying to combine data and schema evolution steps in research data management with provenance aspects [3] and to automatically detect a minimal subset of the original research data being “witnesses” for the research result published afterwards [2]. There are other fundamental research problems such as

- overcoming the heterogeneity of the data,
- deriving a data description, metadata, or database schema that does not exist or is incomplete,
- ensuring the provenance of research results and the reproducibility of scientific evaluations,
- the specification and tracking of scientific workflows, both in terms of organisation and data technology,
- the embedding and storage of application-specific functions and methods, especially for data analysis,
- temporal aspects for the reproducibility of evaluations of measurement data, which are constantly produced from sensors as streaming data,
- as well as the complexity in the evaluations and changes in the evaluation routines over a longer period of time.

While it is important to further examining the theory behind replicability and reproducibility of scientific workflows, the interoperability of heterogeneous data, and the workflow and data provenance problems, it is also important to provide pragmatic and infrastructural solutions and process support for research data management in a timely manner. Pragmatic and infrastructural aspects of research data management are

- Open Science, including FAIR access to and presentation of data and results,
- the feasibility of solutions in practice, for example through flexible architectures,
- the sustainability of the implemented solutions,
- usability or ergonomics of the software system for the researchers that are not IT experts,
- as well as licensing and legal questions regarding original data used and software tools for the evaluation and presentation of the data.

Other approaches also support collaborative work and focus on the collection, management and use of research metadata. Furthermore, research data and (database-supported) evaluations can also be integrated into the document to be published as in Janiform with the Portable DataBase Files (PDbF) [5].

This special issue will present four articles spanning the range from organizational aspects and data management plans in the early phase of the grant submission process to extending the FAIRness principle to the underlying research infrastructure. Additionally, there is a diversity in the data-driven sciences ranging from a deeper FAIR principle by a semantics-aware data analysis in Mathematics to a research data management platform supporting the work of interdisciplinary research teams consisting of experts in Biomedicine, Engineering, and Computer Science.

Submissions to this special issue have been collected from July up to midth of October 2019, the review process was closed midth of November 2019, and the revised articles have been expected in December 2019. The final decision on the revised versions was communicated in January 2020.

3 Overview of the articles in this special issue

For this special issue, we have been able to accept four papers representing different applications and different phases of research data management.

3.1 (Deep) FAIR mathematics

The first article by Katja Berčič, Michael Kohlhase, and Florian Rabe on *(Deep) FAIR Mathematics* provides insights into a deeper semantics for mathematical research data. Modern Mathematics is becoming increasingly data-driven, including both human-curated as well as machine-produced data. A future research infrastructure in the Mathematics community will be open and freely available, consisting of both types of data and the software producing the data. One of the main problems with the mathematical datasets is their diversity and complexity.

Today, in Mathematics the FAIR principle is not supported in most cases. Freely accessible datasets are hard or impossible to reuse, because of the missing metadata annotating the raw research data. Therefore, the authors introduce *deep FAIRness* for mathematical research data. Deep FAIRness extends the classical FAIRness in two aspects:

- The data should become semantics-aware, so the mathematical meaning of the data should be FAIR in all its depth to make the data really interoperable and reusable.
- Not only a complete dataset, but each object or record in the dataset should be identifiable, therefore findable and accessible in a fine-grained manner.

The authors introduce a mathematical data description language (MDDL) and a portal for deep FAIR tabular mathematical data (called DataMathHub).

3.2 Intra-consortia data sharing platforms for interdisciplinary collaborative research projects

The second article by Max Schröder, Hayley LeBlanc, Sascha Spors, and Frank Krüger on *Intra-consortia Data Sharing Platforms for Interdisciplinary Collaborative Research Projects* is pointing out that existing research data management platforms often focus on the publication and archiving of research data at the end of a research project. More important would be the management of research data during research projects, supporting every step in the research process. A good support for this kind of research data management is hard to achieve especially in large interdiscipilinary research consortia.

As an example, the authors present their own research work in the German DFG-funded Collaborative Research Centre (Sonderforschungsbereich; SFB) 1270 ELAINE with researchers from Biomedicine, Engineering, and Computer Science. They show how the CKAN platform – an open source data portal platform for open science data – can be extended in order to serve as an intra-consortia data sharing platform. This is leading to the ELAINE DataHub.

General requirements for such a platform are – besides data sharing – data versioning, data provenance, and interoperability with existing research software and computing infrastructures. Project-specific requirements in ELAINE are the support for simulation as well as wet-lab experiments that are strongly dependent on each other and deeply interconnected, and the handling of very heterogeneous types of data.

3.3 Exploring research data management planning challenges in practice

The third article by Armel Lefebvre, Baharak Bakhtiari, and Marco Spruit on *Exploring Research Data Management Planning Challenges in Practice* presents the organizational and technological challenges occuring during the planning phase of research data management tasks. The planning phase is part of the grant submission process, so a very early phase in the scientific workflow before starting the actual research work. Most of the funding agencies have implemented policies to achieve that the funded projects yield high-quality and reusable research results. In many cases, the researchers are asked to develop data management plans.

The authors of this article investigate current research data management planning practices in academia from two perspectives, a funder perspective and a research data service perspective:

- For the first perspective, they collect experiences from representatives of public funding agencies, grant reviewers, and data stewards.
- For the second perspective, they analyze the data management sections in draft proposals of projects submitted to the Dutch national science foundation.

One specific goal of this examination was to investigate, whether the ambition to produce reusable research data is already reflected in the current research proposals.

3.4 From FAIR research data toward FAIR and open research software

The fourth article by Wilhelm Hasselbring, Leslie Carr, Simon Hettrick, Heather Packer, and Thanassis Tiropanis on *From FAIR Research Data toward FAIR and Open Research Software* is pointing out, that not only research data have to be FAIR, but also the research software implementing the processing steps on these data. For good scientific practices, this software should be open and FAIR, too. Only then, the repeatability, reproducibility, and reuse of research data will be achieved.

The authors analyze the current state in this area to give recommendations for making research software FAIR and open. They consider a pragmatic view and an infrastructure view in this paper. The pragmatic view regards Open Science as a method to make research more efficient by opening the scientific value chain, including external knowledge and allowing collaboration through online tools. The infrastructure view is concerned with software tools, applications, and computing systems.

Among others, they describe artifact evaluation as a review mechanism. Several ACM conferences initiated this, where artifacts can be software systems, scripts, or datasets. For example, the SIGMOD conference of the database research community (Special Interest Group on Management Of Data) calls this special review mechanism *reproducibility evaluation*, where the software and scripts have to be freely accessible and interoperable for the reviewers. To allow an efficient artifact evaluation process, the software developed or used – and not only the research data behind that – has to be FAIR.

4 Summary

The contributions in this special issue provide an overview of different aspects of research data management, ranging from organizational aspects and data management plans in the early phase of the grant submission process to extending the FAIRness principle to the underlying research infrastructure. Additionally, there is a diversity in the data-driven sciences ranging from a deeper understanding of the FAIR principle by a semantics-aware data analysis in Mathematics to a research data management platform supporting the work of interdisciplinary research teams consisting of experts in biomedicine, engineering, and computer science.

Acknowledgement: As the guest editor of this issue, I would like to thank all the authors submitting and revising their contributions within the given deadlines, and all the reviewers for their detailed and constructive comments. Additionally, I would like to thank Stefan Conrad and the associate Editor of this issue, Richard Lenz, especially for their support in using the software infrastructure behind this small part (the submission and review process) of research data management.

References

1. Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. A comparison of research data management platforms: architecture, flexible metadata and interoperability, *Universal Access in the Information Society*, Vol. 16(4), pp. 851–862, 2017.
2. Tanja Auge and Andreas Heuer. The Theory behind Minimizing Research Data — Result equivalent CHASE-inverse Mappings. In *LWDA 2018, CEUR Workshop Proceedings, Vol. 2191*. CEUR-WS.org, pp. 1–12, 2018.
3. Tanja Auge and Andreas Heuer. *Combining Provenance Management and Schema Evolution*. IPAW, Lecture Notes in Computer Science, Vol. 11017, pp. 222–225, Springer, 2018.
4. Ronald F. Boisvert. Incentivizing reproducibility. *Commun. ACM*, Vol. 59(10), pp. 5, 2016.
5. Jens Dittrich and Patrick Bender. Janiform intra-document analytics for reproducible research. *PVLDB*, Vol. 8(12), pp. 1972–1975, 2015.
6. Dror Feitelson. From Repeatability to Reproducibility and Corroboration. *Operating Systems Review*, Vol. 49(1), pp. 3–11, 2015.
7. Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? What form? What from?. *VLDB J.*, Vol. 26(6), pp. 881–906, 2017.
8. Max-Planck-Gesellschaft. *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, 2003. https://openaccess.mpg.de/.

9. Luc Moreau and Paul T. Groth. *Provenance: An Introduction to PROV*. Morgan & Claypool, 2013.
10. Andreas Ruscheinski, Dragana Gjorgevikj, Marcus Dombrowsky, Kai Budde, and Adelinde M. Uhrmacher. *Towards a PROV Ontology for Simulation Models*. IPAW, Lecture Notes in Computer Science, Vol. 11017, pp. 192–195, Springer, 2018.
11. Mark D. Wilkinson, Michel Dumontier, Ijsbrand Jan Aalbersberg, and others. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, Vol. 3, Article No. 160018, 2016.

Bionotes

Prof. Dr. Andreas Heuer
University of Rostock, Computer Science Institute, 18059 Rostock, Germany
andreas.heuer@uni-rostock.de

Prof. Dr. Andreas Heuer studied Mathematics and Computer Science at the Technical University of Clausthal from 1978 to 1984. He got his PhD and Habilitation at the TU Clausthal in 1988 and 1993, resp. Since 1994, he is full professor for Database and Information Systems at the University of Rostock. Andreas Heuer is interested in fundamentals of database models and languages, object-oriented databases and digital libraries, in database support for assistive systems, and in big data analytics, here especially in the four “P”: performance, privacy, preservation, and provenance. One of the main application areas of his research in provenance, preservation, and privacy is research data management.